**FIAP – Faculdade de Informática e Administração Paulista**

Pós Tech - Tech Challenge - Fase 2 – 9DAT

"Utilização de Regressão Logística para prever o fechamento do Ibovespa"

Aline Korb, Gabriel Manzini, Gabriel Teles, Thiago Temporim

São Paulo, SP - 2025

Sumário

[2 Introdução 3](#_Toc204715232)

[3 Escolha do modelo, análise e exploração dos dados 3](#_Toc204715233)

[4 Engenharia de atributos e preparação da base para previsão 3](#_Toc204715234)

[5 Resultados e análises das métricas 4](#_Toc204715235)

[6 Conclusão 5](#_Toc204715236)

# Introdução

O **Ibovespa (Índice Bovespa)** é o principal indicador de desempenho do mercado de ações brasileiro, refletindo a performance das ações mais negociadas na B3 (Bolsa de Valores do Brasil). Ele é amplamente utilizado por investidores, analistas e gestores de fundos como referência para o comportamento do mercado acionário nacional.

Neste relatório, desenvolvemos um modelo de **classificação binária** com base em **regressão logística**, com o objetivo de prever se o Ibovespa fechará em **alta** ou **baixa** no dia seguinte. Essa previsão será feita com base em **dados históricos diários do próprio índice**, considerando variações de preço e outras informações derivadas do comportamento passado.

Para realizar a previsão, foram coletados em torno de **quatro anos de dados históricos** do Ibovespa, disponibilizados publicamente no portal Investing.com, seguidos por um processo de **limpeza, transformação e análise exploratória** antes da modelagem preditiva.

# Escolha do modelo, análise e exploração dos dados

Escolhemos a regressão logística porque ela nos permite medir a probabilidade de alta ou baixa do Ibovespa, em vez de prever valores absolutos. É um modelo robusto, rápido, interpretável e eficiente para capturar a influência de variáveis econômicas sobre a tendência do índice, dando mais confiança na tomada de decisão.

# Engenharia de atributos e preparação da base para previsão

Para o desenvolvimento deste modelo, foram criadas novas colunas no DataFrame com o objetivo de enriquecer a análise da série temporal. Esses indicadores técnicos são amplamente utilizados no mercado financeiro para captar tendências, níveis de volatilidade e possíveis reversões de movimento do índice.

As colunas de Média Móvel de 5 e 10 dias representam, respectivamente, a média dos preços de fechamento do Ibovespa nos últimos 5 e 10 pregões. Elas ajudam a identificar a direção predominante do mercado no curto prazo: uma média móvel crescente pode indicar uma tendência de alta, enquanto uma média decrescente pode sinalizar uma tendência de baixa. Complementando essa análise, o indicador MACD (*Moving Average Convergence Divergence*) representa a diferença entre duas médias móveis exponenciais e serve para captar mudanças de tendência. Sua linha de sinal, chamada MACD *Signal*, é uma média do próprio MACD e funciona como um gatilho para identificar possíveis momentos de compra ou venda, especialmente quando há cruzamento entre essas duas curvas.

A coluna RSI foi criada para usar o indicador de *momentum* de mesmo nome (*Relative Strength Index*) que mede a velocidade e a mudança dos movimentos de preço. O RSI varia de 0 a 100 e é usado para identificar condições de sobrecompra (acima de 70) ou sobrevenda (abaixo de 30), o que pode sugerir uma possível reversão de tendência.

A coluna Vol 5 é a desvio padrão, e representa a volatilidade do mercado nos últimos 5 pregões. Um desvio padrão elevado indica maior oscilação nos preços, o que pode sugerir maior risco e incerteza no curto prazo.

O conjunto de dados selecionado possui 966 pregões (dias úteis onde tiveram negociação na bolsa de valores), onde 936 foram utilizados para treino e 30 para realizarmos os testes no modelo. Com isso, estipulamos que o nosso target é prever diariamente se o fechamento da cotação foi maior que o do dia anterior.

# Resultados e análises das métricas

Com isso, obtivemos um resultado satisfatório de 86.67% de acurácia, abaixo, temos as métricas de avaliação do sklearn:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classe** | **Precision** | **Recall** | **F1-Score** | **Suporte** |
| **0** | 0.88 | 0.88 | 0.88 | 16 |
| **1** | 0.86 | 0.86 | 0.86 | 14 |
| **Accuracy** | — | — | **0.87** | 30 |
| **Macro Avg** | 0.87 | 0.87 | 0.87 | 30 |
| **Weighted Avg** | 0.87 | 0.87 | 0.87 | 30 |

O desempenho foi equilibrado entre as duas classes (alta e baixa), com a classe 0 apresentando ligeiramente maior precisão/recall (88%) e a classe 1 com precisão/recall de 86%. O *f1-score* ficou em 0.88 e 0.86, respectivamente. Esses valores indicam uma performance sólida e consistente do modelo. Para ilustrar melhor os resultados, utilizamos a matriz de confusão que apresenta os acertos e erros sobre a massa de testes:

Gráfico

O conteúdo gerado por IA pode estar incorreto.

Para chegarmos nos resultados finais, passamos por alguns desafios onde a preparação dos dados foi essencial para superamos essa etapa, usamos médias móveis como forma de resumir o comportamento recente dos dados ao longo do tempo, transformando dependências temporais em variáveis numéricas que o modelo consegue entender.

Durante a seleção do modelo de regressão logística, realizamos testes com diferentes tipos de regularização (L1 e L2) e valores do hiperparâmetro C, que controla o grau de penalização sobre os coeficientes do modelo. O melhor modelo foi com regularização L1 e C=1, atingindo: 89.36% de acurácia na validação 86.67% no teste final (últimos 30 dias).

|  |  |  |  |
| --- | --- | --- | --- |
| **Penalty** | **C** | **Acurácia Treino (%)** | **Acurácia Validação (%)** |
| L1 | 0.01 | 50.53 | 50.00 |
| L1 | 0.1 | 85.83 | 88.83 |
| L1 | 1 | 88.10 | 89.36 |
| L1 | 10 | 88.50 | 87.23 |
| L2 | 0.01 | 69.52 | 71.28 |
| L2 | 0.1 | 75.80 | 81.38 |
| L2 | 1 | 85.16 | 88.83 |
| L2 | 10 | 87.70 | 87.77 |

Essa pequena queda de desempenho indica que o modelo não sofreu overfitting relevante e ficou bem generalizado. A escolha da regularização L1 ajudou a reduzir a complexidade do modelo, focando apenas nas variáveis mais relevantes.

# Conclusão

O modelo desenvolvido demonstrou ser eficaz na previsão da direção do fechamento do Ibovespa, alcançando uma acurácia final de 86,67% nos dados de teste. A escolha da regressão logística como método preditivo se mostrou apropriada para o objetivo final, principalmente pela capacidade de lidar com variáveis financeiras derivadas de séries temporais.

A construção do modelo envolveu um processo cuidadoso de coleta, limpeza e preparação dos dados, além da criação de variáveis derivadas do mercado financeiro, como médias móveis, MACD, RSI e volatilidade, que enriqueceram a base e permitiram ao modelo capturar padrões relevantes no comportamento do mercado. Além disso, a utilização da regularização L1 contribuiu para a seleção automática de variáveis mais informativas, promovendo simplicidade sem perda significativa de desempenho.

Apesar da queda leve entre a validação e o teste, os resultados foram consistentes e indicam boa capacidade de generalização. A utilização desse modelo é indicado para sistemas de apoio à decisão em investimentos, com potencial de aprimoramento com o uso de mais dados, variáveis macroeconômicas e teste de abordagens mais complexas, como modelos baseados em redes neurais.

Por fim, este trabalho reforça o potencial do uso de modelos estatísticos e aprendizado de máquina na análise de mercados financeiros, desde que acompanhados de uma preparação cuidadosa dos dados e validações rigorosas.